EESP Workshop SC26 - Chicago, 15 November 2026

# A Deployment-Aware Multi-Objective Hyper-Parameter Optimization Framework for Machine Learning

**Matteo Zanotto**[1], **Romeo Kienzler**[2], **Gabriele Padovani**[1], **Sandro Fiore**[1]

[1]University of Trento (Italy), [2]IBM Research (Switzerland)

Funded by NextGenerationEU
Grant Agreement n. 101086248 - CLOUDSTARS

# Motivation

➔ Sharp rise in computational requirements and energy consumption of modern AI models featuring million to billion params

# Motivation

- ➔ Sharp rise in computational requirements and energy consumption of modern AI models featuring million to billion params

- ➔ Model compression techniques (pruning, quantization) and device-specific optimized runtimes (TensorRT, vLLM)
- ➔ HPO techniques for optimal configuration searching through Bayesian Optimization

# Motivation

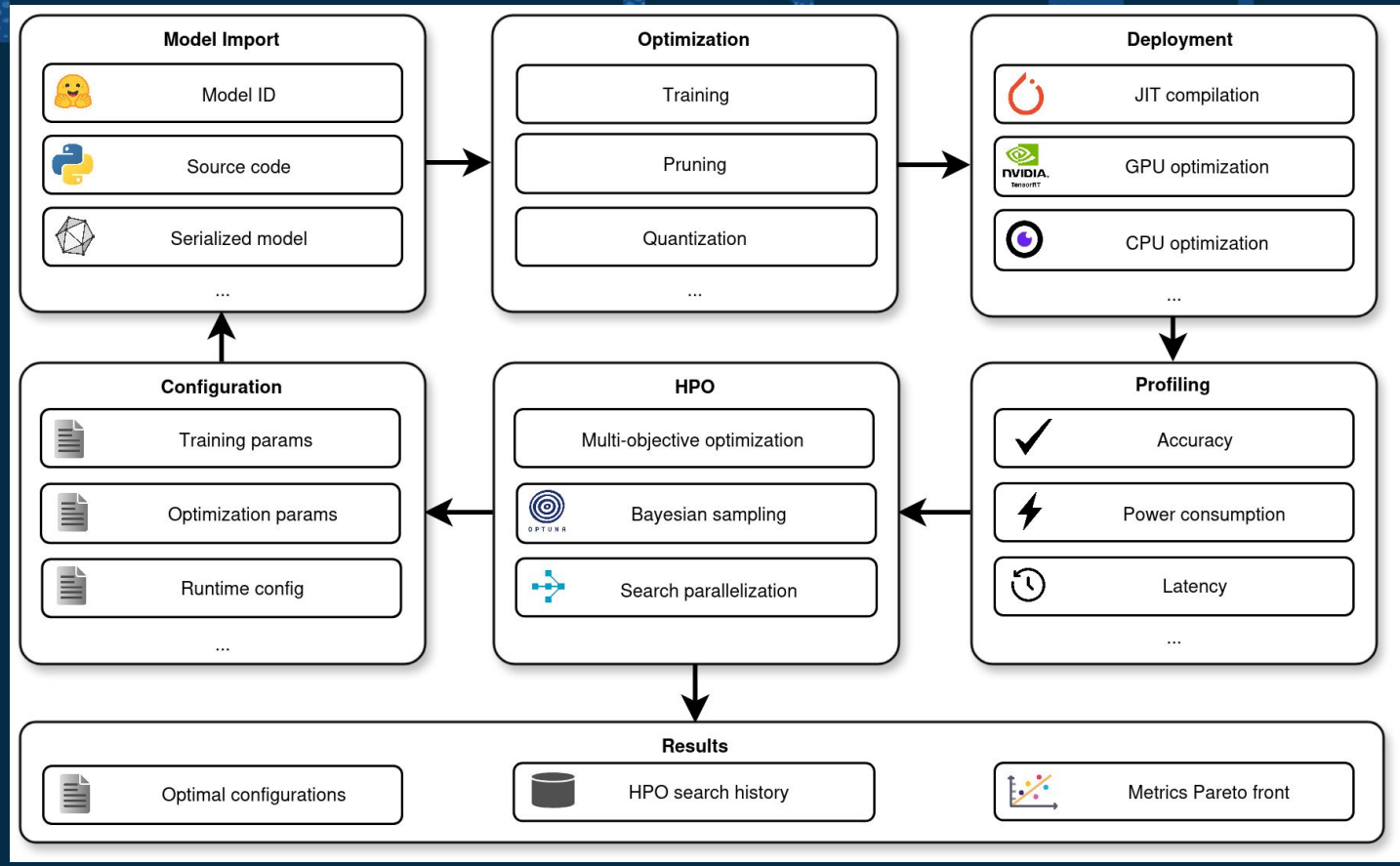
- ➔ Sharp rise in computational requirements and energy consumption of modern AI models featuring million to billion params

- ➔ Model compression techniques (pruning, quantization) and device-specific optimized runtimes (TensorRT, vLLM)
- ➔ HPO techniques for optimal configuration searching through Bayesian Optimization

- ➔ Existing research primarily focused on model development and training time
- ➔ Need for manual coordination of model deployment optimizations

# Methodology

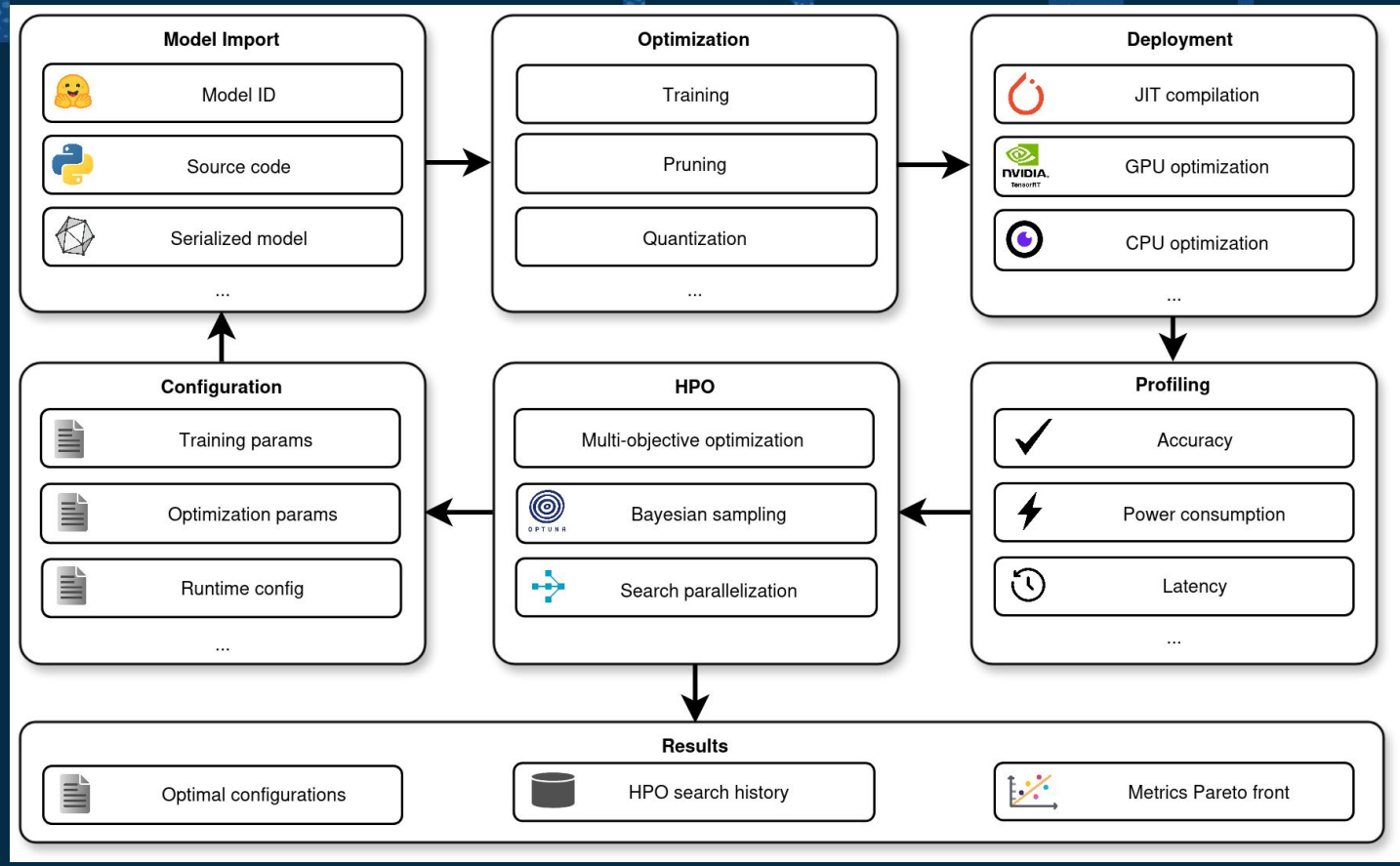
Multi-objective HPO to optimize model deployment



# Methodology

Multi-objective HPO to optimize model deployment

End-to-end MLOps pipeline: from model import to inference config

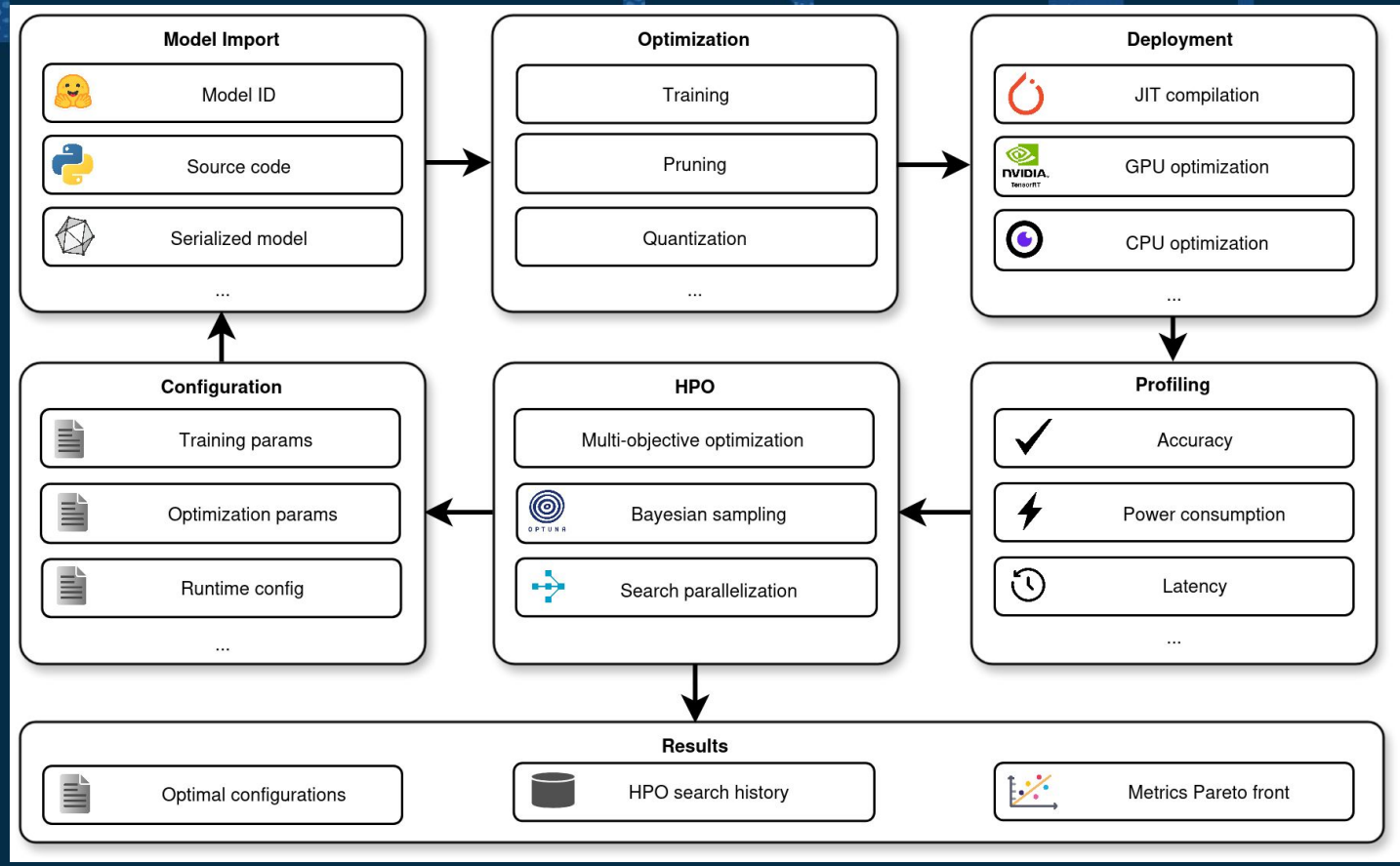


# Methodology

Multi-objective HPO to optimize model deployment

End-to-end MLOps pipeline: from model import to inference config

Iterate*: orchestrate Bayesian Search through Optuna, parallelization with Ray




* https://github.com/claimed-framework/claimed#terratorch-iterate

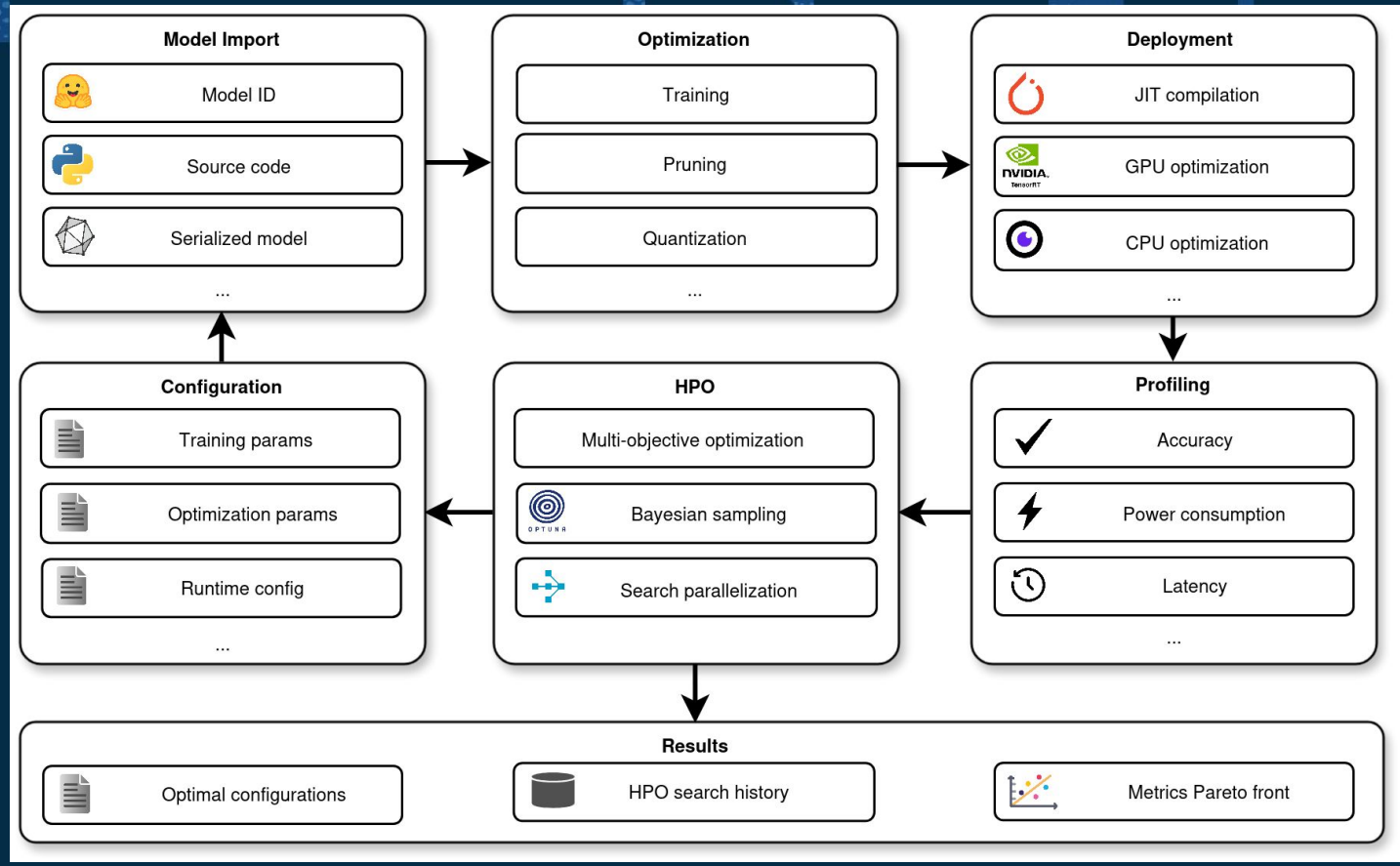
# Methodology

Multi-objective HPO to optimize model deployment

End-to-end MLOps pipeline: from model import to inference config

Iterate*: orchestrate Bayesian Search through Optuna, parallelization with Ray

Pareto front of optimal configurations across objectives




* https://github.com/claimed-framework/claimed#terratorch-iterate

# Evaluation: computer vision on the edge

**Model:** YOLOv (object detection)
**Target:** NVIDIA Jetson Nano

# Evaluation: computer vision on the edge

**Model:** YOLOv (object detection)
**Target:** NVIDIA Jetson Nano

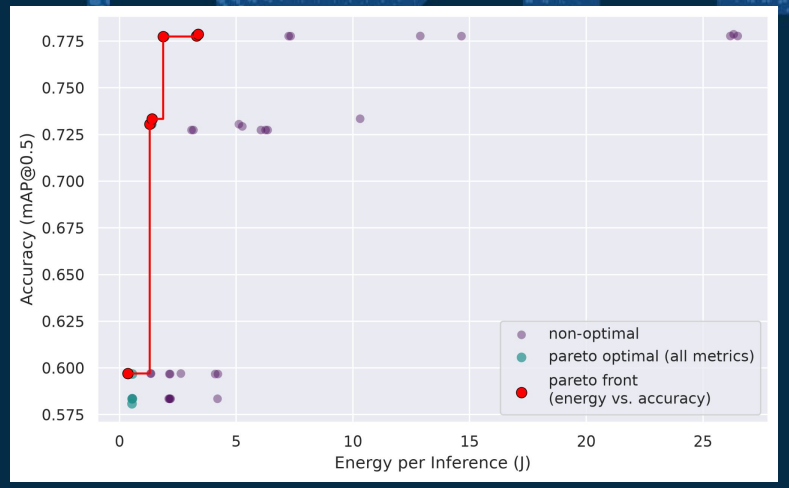
**Objectives:** throughput, latency, accuracy, energy
**Params:** model size, quantization, batch size, tactic, workspace size

# Evaluation: computer vision on the edge

**Model:** YOLOv (object detection)
**Target:** NVIDIA Jetson Nano

**Objectives:** throughput, latency, accuracy, energy
**Params:** model size, quantization, batch size, tactic, workspace size

14 optimal configs
-89% energy ➜ -23% accuracy
+18% accuracy ➜ +800% energy






# Evaluation: LLM serving

**Model:** Qwen 2.5
**Target:** NVIDIA A100 SXM

# Evaluation: LLM serving

**Model:** Qwen 2.5
**Target:** NVIDIA A100 SXM

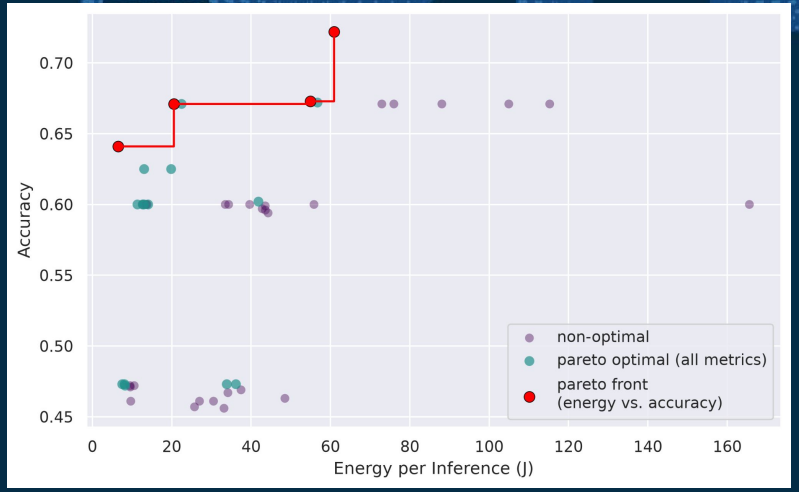
**Objectives:** latency, accuracy, energy, VRAM usage
**Params:** model size, quantization, context length, runtime (vLLM, ONNX, torch compile, python)

# Evaluation: LLM serving

**Model:** Qwen 2.5
**Target:** NVIDIA A100 SXM

**Objectives:** latency, accuracy, energy, VRAM usage
**Params:** model size, quantization, context length, runtime (vLLM, ONNX, torch compile, python)

19 optimal configs
-90% energy ➜ -11% accuracy
+8% accuracy ➜ +200% energy






# Evaluation: HPC benchmark experiment

**Model:** CNN-LSTM (ClimateBench)
**Target:** Frontier (AMD Instinct MI250X)

# Evaluation: HPC benchmark experiment

**Model:** CNN-LSTM (ClimateBench)
**Target:** Frontier (AMD Instinct MI250X)

**Objectives:** inference break-even point, parallel HPO vs. parallel training
**Params:** model architecture, learning rate, batch size, precision

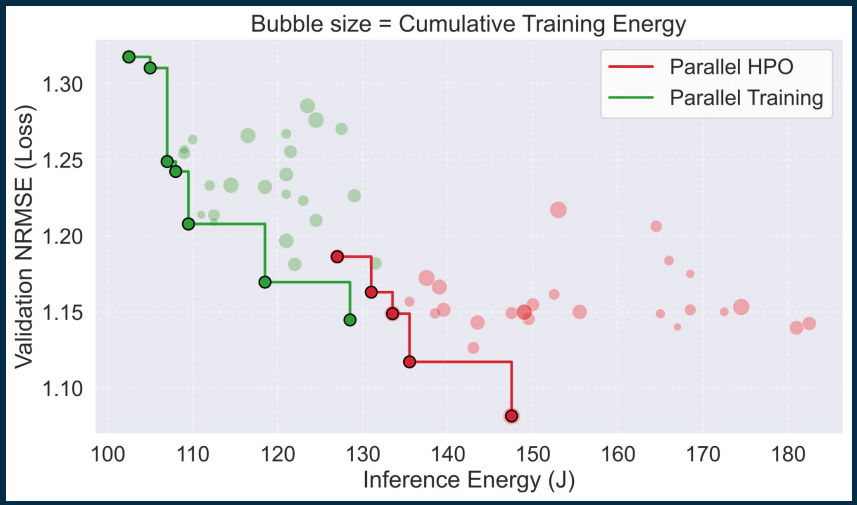
# Evaluation: HPC benchmark experiment

**Model:** CNN-LSTM (ClimateBench)
**Target:** Frontier (AMD Instinct MI250X)

**Objectives:** inference break-even point, parallel HPO vs. parallel training
**Params:** model architecture, learning rate, batch size, precision

best model trade-off at 2000J energy cost for HPO repaid in ~7000 inference requests

| # iterations | HPO energy cost (J) | Best NRMSE | Break-even inferences |
|---|---|---|---|
| 1 | 227.0 | 1.1402 | 479 |
| 15 | 1790.5 | 1.1172 | 7423 |
| 29 | 3348.5 | 1.0816 | 12912 |



EESP Workshop SC26 - Chicago, 15 November 2026

# A Deployment-Aware Multi-Objective Hyper-Parameter Optimization Framework for Machine Learning

**Matteo Zanotto**[1], **Romeo Kienzler**[2], **Gabriele Padovani**[1], **Sandro Fiore**[1]

[1]University of Trento (Italy), [2]IBM Research (Switzerland)

Funded by NextGenerationEU
Grant Agreement n. 101086248 - CLOUDSTARS

EESP WORKSHOP 2026

SC26 hpc unites | Chicago